ADAPTIVE QUANTIZATION USING A PERCEPTUAL VISIBILITY PREDICTOR

Olivier Verscheure

Telecommunication Services Group
Swiss Federal Institute of Technology
CH-1015 Lausanne, Switzerland
verscheure@tcom.epfl.ch

Christian J. van den Branden Lambrecht

Hewlett-Packard Laboratories
1501 Page Mill Road, MS 1U20
Palo Alto, CA 94304
vdb@hpl.hp.com

ABSTRACT

This work addresses the optimization of video services by improving the bit allocation stage in a video encoder. We first introduce a new block-based video metric based on a vision model, termed the perceptual visibility predictor (PVP), which accounts for spatial as well as temporal perceptual activities. In other words, we propose a way to classify sequence areas in terms of their relevance to human perception. We then briefly show how this metric may be used within a one-pass causal adaptive quantization scheme applied to MPEG-2 video encoding.

1. INTRODUCTION

Audiovisual applications (e.g., video conferencing, video on demand, teleteaching, etc.) are foreseen as one of the major users of broadband networks such as ATM networks. At the heart of this revolution is the digital compression of audio and video signals. The biggest advantage of compression resides in data rate reduction which results in reduction of transmission costs. The choice of the compression algorithm mostly depends on the available bandwidth or storage capacity and the features required by the application. The MPEG-2 [1] standard [1], a truly integrated audio-visual standard developed by the International Organization for Standards (ISO), is capable of compressing NTSC or PAL video into an average bit rate of 3 to 6 Mbits/s with a quality comparable to analog CATV [2].

A lot of work remains to be done to optimize these audiovisual applications so that they can be offered at attractive prices. In other words, the user expects an adequate programme quality at the lowest possible cost. One of the major issues facing this optimization is the introduction of vision science knowledge in the video encoding process. Indeed, it is useless to transmit data that the end-user is unable to notice (perceptual redundancy). High compression ratio may be reached by trying to eliminate the perceptual redundancy as much as possible, in addition to reducing the spatial and temporal redundancies. The perceptual redundancy corresponds to video information still present in the bit stream after spatio-temporal compression but which is not perceived by the viewer. Hence, for a given amount of bits to be allocated to a video sequence, the repartition of these bits among and within the pictures may be very different whether the user perception is considered or not.

The present work is the continuation of the approach proposed in [3]. In the latter paper, we have introduced a local video activity metric (MPAM [2]), accounting for both spatial and temporal perceptual activities, that could be estimated independently of the video encoding process. This work focuses on a more efficient block-based video metric, namely the perceptual visibility predictor (PVP). The paper is organized as follows: Section 2 briefly introduces the vision model, on which the proposed metric relies. Section 3 presents the perceptual visibility predictor metric and experimental results are described in Sec. 4. Concluding remarks are given in Sec. 5.

2. THE VISION MODEL

Several studies have shown that a correct estimation of subjective quality has to incorporate some modeling of the human visual system [4, 5, 6]. The spatio-temporal vision model described in [6] had been used to design a computational quality metric for moving pictures [7] which proved to behave consistently with human judgment. Since the PVP is deduced from this metric, we now shortly introduce its principles.

Basically, the metric, termed Moving Pictures Quality Metric (MPQM), first decomposes an original sequence and a distorted version of it into perceptual components using a Gabor filter bank. The filter bank models the multiresolution architecture of the primary visual cortex. It analyzes the visual information in bands tuned in spatial frequency, temporal frequency and orientation. The next stage models pattern sensitivity (i.e., the dependence of sensitivity with frequency). A modeling of pattern sensitivity is computed accounting for contrast sensitivity as well as visual masking.

This visual masking phenomenon is a destructive interference between various stimuli in a single visual sensation. We modeled masking by a non-linear transducer (see Fig. 2) that relates the detection threshold of a target stimulus, C_T as a function of the contrast of the background onto which it lies, C_M. The actual expression we used is presented in Eq. 1, where C_{T0} denotes the contrast threshold of the

[1] MPEG stands for Moving Picture Experts Group

[2] MPAM stands for Moving Pictures Activity Metric

0-8186-8183-7/97 $10.00 © 1997 IEEE

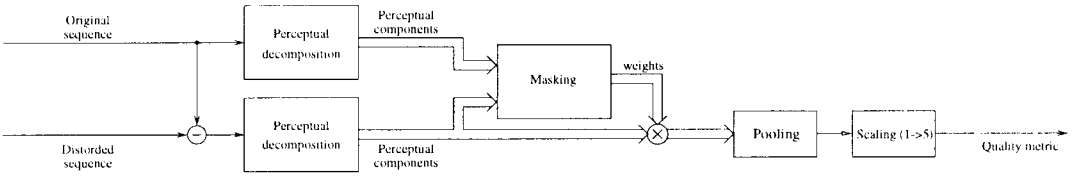

Figure 1: *Block Diagram of the Moving Pictures Quality Metric (MPQM).*

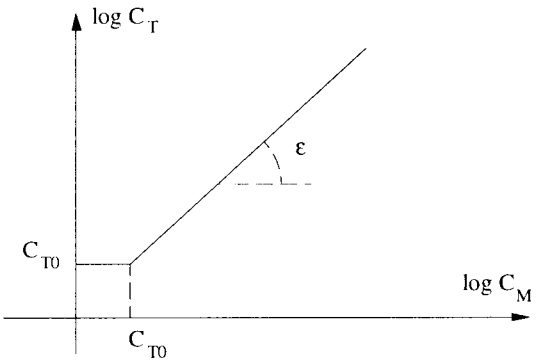

Figure 2: *Model of the masking phenomenon.*

considered stimulus in the absence of a background. This computation is done within each channel and only considers interaction between stimuli belonging to the same channel.

$$C_T = \begin{cases} C_{T0}, & if\ C_M < C_{T0} \\ C_{T0} \cdot [\frac{C_M}{C_{T0}}]^{\epsilon}, & otherwise. \end{cases} \quad (1)$$

Finally, the data is pooled over the channels (using a Minkowski summation) to compute the *local perceived error* which is then scaled using a non-linear function to obtain a quality rating between 1 and 5 according to the ITU-R 500.3 standard [8] (the higher the value, the better the quality). The block diagram of the MPQM is illustrated in Fig. 1).

3. THE PERCEPTUAL VISIBILITY PREDICTOR

The proposed PVP metric aims at predicting the *local perceived error* by using the above concepts. Its block diagram is presented in Fig. 3. The original video sequence is first decomposed into perceptual channels by the three-dimensional filter bank that emulates the detection mechanisms of the cortex. The next stage models pattern sensitivity (i.e., appearance of the stimuli as a function of their frequency). In our model, pattern sensitivity computes the detection threshold for the coding noise due to the scene[3] according to Eq. (1). This is equivalent to computing the local perceptual activity of the scene, on a channel basis.

The last operation, denoted "pooling", simulates higher order integration by subsequent areas of the cortex. It gathers the channels measurements together to yield a single measurement. This measurement is termed the *visibility predictor* since it predicts the perceptual relevance of a region. Let $C_T(a,b,t,c)$ be the computed detection threshold at position (a,b), at time t and in channel c. The Perceptual Visibility Predictor, PVP, for a region of spatial dimension N_x and N_y, centered around (x,y) at time t is obtained using the following Minkowski summation:

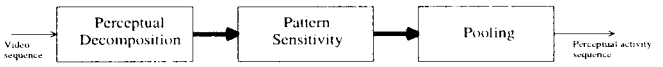

Figure 3: *Perceptual Visibility Predictor (PVP) block diagram. The thick arrows represent a set of perceptual components while the thin ones represent video sequences.*

$$\text{PVP} = \left(\frac{1}{N} \sum_{c=1}^{N} \left(\frac{1}{N_x N_y} \sum_{a=x-\frac{N_x}{2}}^{x+\frac{N_x}{2}} \sum_{b=y-\frac{N_y}{2}}^{y+\frac{N_y}{2}} |C_T[a,b,t,c]| \right)^{\beta} \right)^{\frac{1}{\beta}},$$

where, $N = 34$ is the number of perceptual channels and β has a value of -2. The latter value has been chosen for the following reasons: the higher the C_T value, the stronger the visual masking, the lower the remaining noise after visual masking (in a first approximation). At this stage, it is desirable to have the behavior opposite to the one of MPQM: PVP should increase as visual masking decreases (low C_T values must have stronger influence than the high ones). So, we perform the pooling operation with an exponent that is the symetric of 4 with respect to a value of 1 (which would correspond to a simple averaging).

Compared to the MPAM metric described in [3], the proposed PVP metric only requires a single perceptual decomposition (with a smaller number of channels) and does not require any additional synthetic input, nor any reconstruction stage. However, both metrics give results that correlate well with human judgment.

4. EXPERIMENTAL RESULTS

In our simulations, two 32 picture-long sequences of 512 × 512 pixels have been used ("Mobile&Calendar" and "FlowerGarden"). These sequences have been encoded, as interlaced video, with a structure of 12 images per GOP and 2 B-pictures between every reference picture, in open-loop VBR mode (i.e., a constant quantizer scale value has been used over the whole sequence). For that purpose, a modified version of the TM5 MPEG-2 software encoder [9] has been used.

[3]the scene can be considered as a masker with respect to coding noise as part of the noise is masked by the scene.

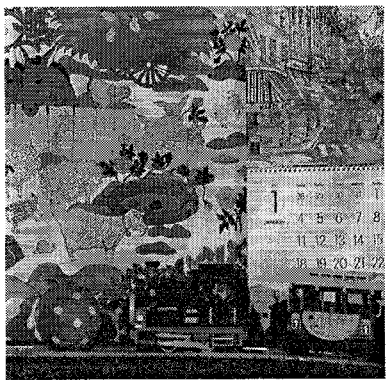

Figure 4: 13th picture of the "Mobile&Calendar" sequence.

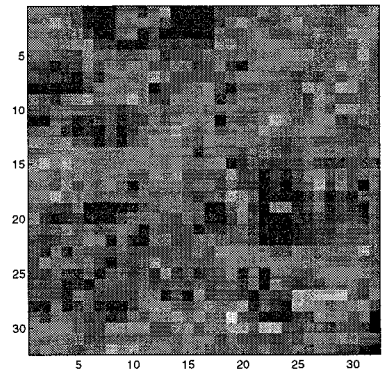

Figure 5: Macroblock-based perceived error map (constant quantizer scale value of 32).

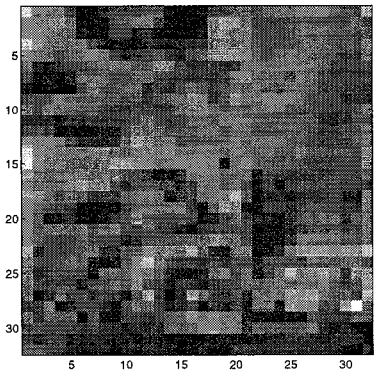

Figure 6: Macroblock-based perceptual visibility predictor (PVP) map.

The block-based PVP values ($N_x = N_y = 8$) have been computed for both sequences. Figure 6 shows the macroblock-based PVP map corresponding to the $13th$ picture (coded as an I-picture) of the "Mobile&Calendar" sequence (see Fig. 4). The PVP value, PVP_i, of a macroblock i has been chosen as the maximum among the four 8×8-block PVP values (a macroblock is not better than the block of highest visible distortion). In a given macroblock, the higher (lighter areas) the PVP value, the higher the annoying impact of the coding noise perceived by the user. Figure 5 represents the macroblock-based perceptual error map (MPEG-2 encoding noise) obtained after open-loop encoding of the Mobile&Calendar sequence with a constant quantizer scale value of 32.

The same simulations have been performed on the "FlowerGarden" sequence. Figures 9 and 8 illustrate, respectively, the macroblock-based PVP and the macroblock-based perceptual error maps corresponding to the $4th$ picture (coded as a P-picture).

From the previous results, it is to be noted that, according to the macroblock-based perceptual error, the PVP metric seems to behave consistently. This conclusion holds for different quantizer scale values. However, the correlation between the two signals can be increased by considering only the PVP values under a given threshold. Indeed, high PVP values seem to lead to unreliable results. This is an issue that we are currently investigating.

At this point, an adaptive perceptual bit allocation, compliant with the MPEG-2 syntax, can be performed. As proposed in [10], we separate the quantizer scale into two multiplicative factors, a nominal quantization term, Q, and a perceptual quantization term, Q_P, so that $Q_S = Q \cdot Q_P$. Q_P is used to compensate for the spatio-temporal characteristics of the block to be quantized so that all blocks quantized with the same value of Q will have the same perceptual quality even though the actual used quantizer scale might be different. The lower the PVP value, the higher the perceptual quantization term, Q_P, in order to produce approximately uniform perceptual coding noise, and conversely. However, since high PVP values may lead to unreliable decisions, the nominal quantization term, Q, is not modulated anymore as soon as the corresponding PVP value reaches a given threshold, T. The perceptual quantization value for the macroblock i, $Q_P(i)$, is then defined as:

$$Q_P(i) = \begin{cases} K \cdot \frac{PVP(i)+2 \cdot PVP_{avg}}{2 \cdot PVP(i)+PVP_{avg}}, & if\ PVP(i) < T \\ 1, & otherwise. \end{cases} \quad (2)$$

where, $\text{PVP}(i)$ is the PVP value associated to the macroblock i, PVP_{avg} is the average PVP value over all the macroblocks in the picture containing the macroblock i and K is a function of Q.

Simulations have been performed on the "FlowerGarden" sequence using a nominal quantization term, Q, equal to 32. The sequence has been encoded in open-loop VBR mode, with a structure of 12 images per GOP and 2 B-pictures between every reference picture, using both a constant quantizer scale value of 32 over the whole sequence and the perceptual bit allocation as described above. Although this is a preliminary study, the integration of the PVP metric in the MPEG-2 bit allocation stage permitted to save up to 9% of the bit rate for the same perceptual video quality. This gain in bit rate for a same video quality is explained by the reduction of the video quality's variance.

5. CONCLUSIONS AND FUTURE WORKS

In this paper, we have presented the perceptual visibility predictor (PVP). The PVP permits to classify sequence areas in terms of their relevance to human perception. Simulations, on different sequences open-loop encoded with different quantizer scale values, have shown a consistent behavior of the PVP. We also have roughly sketched a way to use this metric within a one-pass caudal adaptive quantization scheme applied to MPEG-2 video encoding.

Further works will be carried out in order to elaborate a more efficient bit allocation scheme based on the PVP metric. Once improved, such a scheme will serve as the target for optimization of the rate control mechanism proposed in [11]).

Figure 7: 4th picture of the "Flower-Garden" sequence.

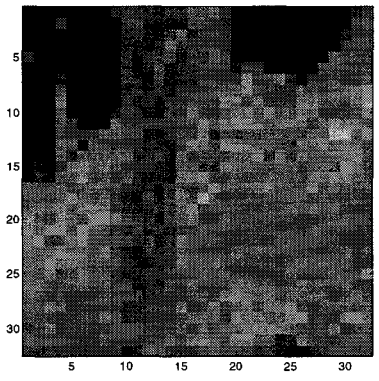

Figure 8: Macroblock-based perceived error map (constant quantizer scale value of 32).

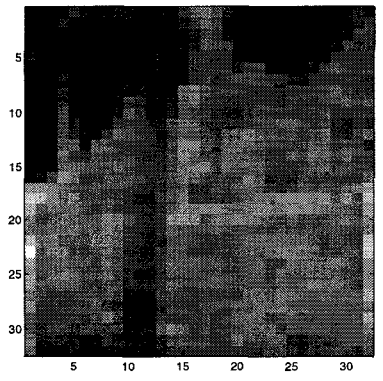

Figure 9: Macroblock-based perceptual visibility predictor (PVP) map.

6. REFERENCES

[1] I. J. 1, *Information Technology - Generic Coding of Moving Pictures and Associated Audio Information - Part 1, 2 and 3.* ISO/IEC JTC 1, 1994.

[2] Barry G. Haskell, Atul Puri and Arun N. Netravali, *Digital Video: an Introduction to MPEG-2.* Digital Multimedia Standards Series, Chapman and Hall, 1997.

[3] Olivier Verscheure, Andrea Basso, Mounir El-Maliki and Jean-Pierre Hubaux, "Perceptual Bit Allocation for MPEG-2 Video Coding," in *Proceedings of the International Conference on Image Processing,* (Lausanne, Switzerland), September 16-19 1996. Available on `http://tcomwww.epfl.ch/~verscheu/`.

[4] S. Comes, *Les traitements perceptifs d'images numerisées.* PhD thesis, Université Catholique de Louvain, 1995.

[5] J. A. Saghri, P. S. Cheatham and A. Habibi, "Image Quality measure based on a Human Visual System Model," in *Optical Engineering,* vol. 28, pp. 813–818, 1989.

[6] C. van den Branden Lambrecht, *Perceptual Models and Architectures for Video Coding Applications.* PhD thesis, Swiss Federal Institute of Technology, CH-1015 Lausanne, Switzerland, 1996. Available on `http://ltswww.epfl.ch/pub_files/vdb/`.

[7] C. J. van den Branden Lambrecht and O. Verscheure, "Perceptual Quality Measure using a Spatio-Temporal Model of the Human Visual System," in *Proceedings of the SPIE,* vol. 2668, (San Jose, CA), pp. 450–461, January 28 - February 2 1996. Available on `http://tcomwww.epfl.ch/~verscheu/`.

[8] ITU-R Commity, "ITU-R 500.3 Recommendations and Reports," in *ITU,* vol. XI, 1986.

[9] Chadd Fogg, "mpeg2encode/mpeg2decode version 1.1. available via anonymous ftp at ftp.netcom.com," in *MPEG Software Simulation Group,* June 1994.

[10] Dzung T. Hoang, Elliot Linzer and Jeffrey S. Vitter, "Lexicographically Optimal Rate Control for Video Coding with MPEG Buffer Constraints," tech. rep., Duke University Computer Science, February 1996.

[11] Olivier Verscheure, Philippe Chabloz and Jean-Pierre Hubaux, "A Flexible Rate Control Mechanism for Interactive MPEG-2 Video Services over ATM Networks," in *International Workshop on Audio-Visual Services over Packet Networks,* (Aberdeen, Scotland, UK), September 1997. Available on `http://tcomwww.epfl.ch/~verscheu/`.